Organizing Videos Streams for Clustering and Estimation of Popular Scenes

Battiato S[1], Farinella GM[1], Milotta FLM[1,2], Ortis A[1,2], Stanco F[1], D'Amico V[2], Addesso L[2], Torrisi G[2]
1 - University of Catania (Italy), 2 – Telecom Italia JOL WAVE (Italy)
{battiato, gfarinella, milotta, ortis, fstanco}@dmi.unict.it, {valeria1.damico, luca.addesso, giovanni.torrisi}@telecomitalia.it

19th International Conference on Image Analysis and Processing
11—15 September 2017
Catania — Italy

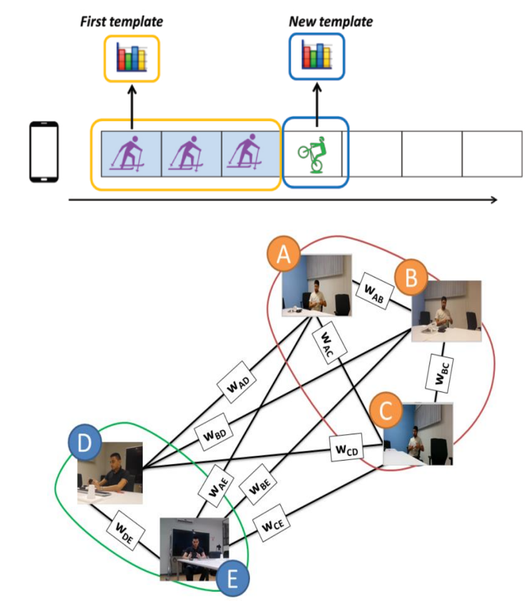
Abstract

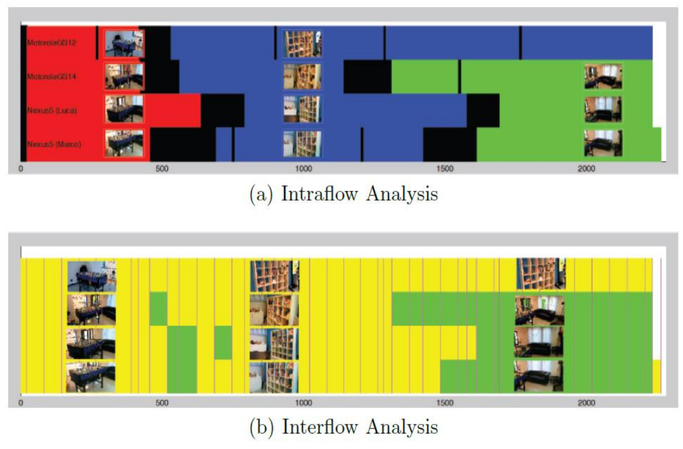
The huge diffusion of mobile devices with embedded cameras has opened new challenges in the context of the **automatic understanding of video streams acquired by multiple users** during events, such as sport matches, expos, concerts. Among the other goals there is the interpretation of which visual contents are the most relevant and popular (i.e., where users look). The *popularity of a visual content* is an important cue exploitable in several fields that include the estimation of the mood of the crowds attending to an event, the estimation of the interest of parts of a cultural heritage, etc. In live social events people capture and share videos which are related to the event. The popularity of a visual content can be obtained through the ***"visual consensus"*** among multiple video streams acquired by the different users devices. In this paper we address the problem of detecting and summarizing the *"popular scenes"* captured by users with a mobile camera during events. For this purpose, we have developed a framework called **RECfusion** in which the key popular scenes of multiple streams are identified over time. The proposed system is able to **generate a video which captures the interests of the crowd starting from a set of the videos by considering scene content popularity**. Frames composing the final popular video are automatically selected from the different video streams by considering the scene recorded by the highest number of users' devices (i.e., the most popular scene).

Dataset is available at: http://iplab.dmi.unict.it/recfusionICIAP17

Pipeline

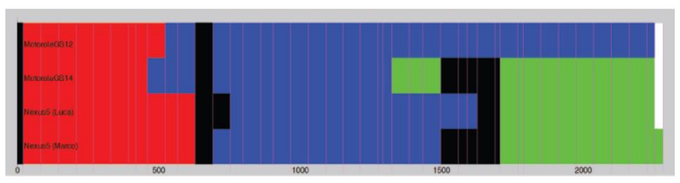
1. Video acquisition phase (multiple sources)
2. **Intraflow Analysis**
3. **Interflow Analysis**
4. **Cluster Tracking**
5. Automatic generation of final video by considering *scene content popularity*





(a) Intraflow Analysis

(b) Interflow Analysis



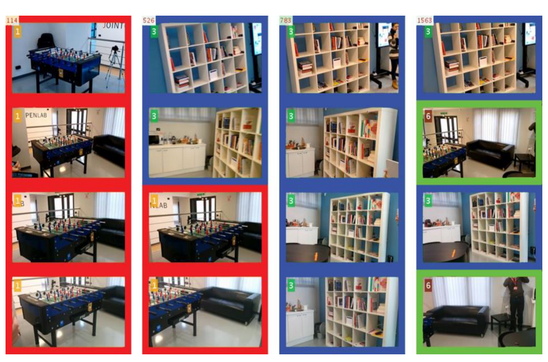
(c) Cluster Tracking

RECfusion Dataset

1. ***Foosball***: 4 devices, ~2250 frames
2. ***Meeting***: 5 devices, ~2895 frames
3. ***S.Agata***: 7 devices, ~1258 frames

Additional Video Dataset employed:

4. ***Magician [1]***: 6 devices, ~3800 frames
5. ***Hoshen [2]***: 3 video sets



Experimental Results

Table 1. Validation Results of Popularity Estimation.

Scenario	Devices	Models	P_a/P_r	P_g/P_r
Foosball	4	2	1.02	1
Meeting	2	2	1.01	0.99
Meeting	4	4	0.99	0.95
Meeting	5	5	0.89	0.76
SAgata	7	6	1.05	1
Magician	6	6	0.73	0.73
Concert [5]	3	1	1.06	1
Lecture [5]	3	1	1.05	0.86
Seminar [5]	3	1	0.62	0.62

The ratio P_a/P_r provides a score for the **popularity estimation**, whereas the ratio P_g/P_r verifies the visual content of the videos in the popular cluster and provides a measure of the **quality of the popular cluster**. Note that P_a/P_r is a score: when is lower than 1 it means that system is under-estimating the popularity of the cluster, while, conversely, if it is higher than 1 it results in an over-estimation.

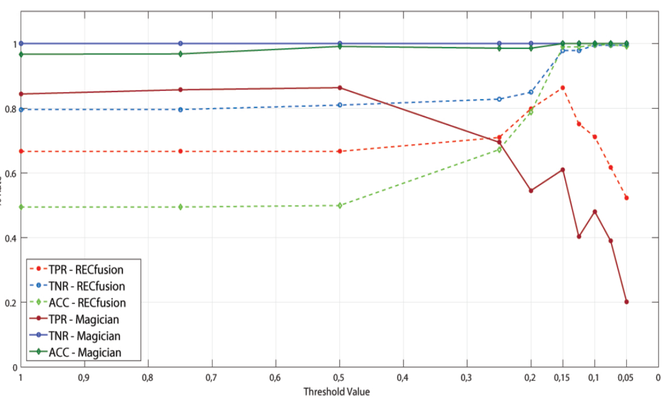
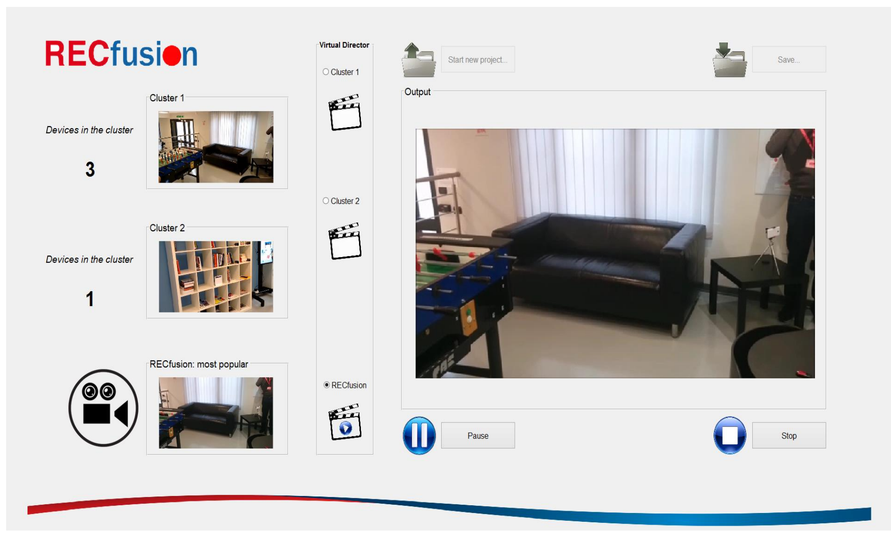


Table 2. Validation results between cluster tracking procedure threshold-based and vote-based.

DS	Scene	TPR (RECALL) [3]	TPR (RECALL) PROPOSED	TNR (SPECIFICITY) [3]	TNR (SPECIFICITY) PROPOSED	ACC (ACCURACY) [3]	ACC (ACCURACY) PROPOSED
Foosball	1		**0,92**		**1,00**		**1,00**
	2	0,69	**0,97**	**0,98**	0,91	**0,99**	0,97
	3	0,41	**0,74**	**1,00**	**1,00**	0,50	**1,00**
	MEAN	0,67	**0,87**	0,89	**0,97**	0,73	**0,99**
Meeting	1	0,99	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
	2	0,80	**1,00**	**0,95**	0,93	**0,83**	0,67
	3	0,43	**0,50**	**1,00**	**1,00**	0,70	**1,00**
	MEAN	0,74	**0,83**	**0,98**	**0,98**	0,84	**0,89**
S.Agata	1	0,71	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
	2	0,87	**0,97**	**0,49**	0,14	**0,80**	0,68
	3	**0,48**	0,00	**1,00**	**1,00**	**0,60**	0,00
	MEAN	**0,69**	0,66	**0,83**	0,71	**0,80**	0,56
Magician	1	0,73	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**
	2	0,45	**0,56**	**1,00**	**1,00**	**0,98**	0,91
	MEAN	0,59	**0,78**	**1,00**	**1,00**	**0,99**	0,96

RECfusion GUI



References

[1] L. Ballan, G.J. Brostow, J. Puwein, and M. Pollefeys. Unstructured video-based rendering: Interactive exploration of casually captured videos. In ACM Transactions on Graphics, pages 1-11, 2010.

[2] Y. Hoshen, G. Ben-Artzi, and S. Peleg. Wisdom of the crowd in egocentric video curation. In IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 587-593, 2014.

[3] F.L.M. Milotta, S. Battiato, F. Stanco, V. D'Amico, G. Torrisi, and Luca Addesso. Recfusion: Automatic scene clustering and tracking in video from multiple sources. In EI - Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications, 2016.

Conclusion

In this paper we described **RECfusion**, a framework designed for automatic video curation driven by the popularity of the scenes acquired by multiple devices. Given a set of video streams as input, the framework can group these video streams by means of *similarity* and *popularity*, then it **automatically suggests a video stream to be used as output**, acting like a "*virtual director*". We compared RECfusion *intraflow* and *interflow analysis* validations with Hoshen ***[2]***. We have added a video set from Ballan ***[1]*** to our RECfusion dataset showing that RECfusion is capable to recognize and track the scenes of a video collection even if there is a single scene, where all the user are focused on the same target and videos are affected by severe camera motion. We proposed a novel and alternative **vote-based cluster tracking procedure** and compared it with the one, threshold-based, described in ***[3]***. From this comparison we found that **vote-based procedure reaches very good results totally automatic and independently by a hyperparameter fine tuning phase**, but with the tradeoff of be unable to create and track an unlimited number of clusters. As future works and possible applications, we are planning to augment the framework with features specifically focused on Assistive Technology or Security issues (i.e., highlight/track bad behavior in the life style, log the visited places, search something or someone that appears in the scene).